FDA Submission

Your Name: Mayank Sharma

Name of your Device: Pneumonia Detector Ultra Max Pro

Algorithm Description

1. General Information

Intended Use Statement: Assisting radiologists in detection of pneumonia in x-ray images.

Indications for Use:

- Screening of Pneumonia Studies in Chest X-Ray.
- This algorithm is intended for use on men and women from 1 to 75 years old who have no previous illnesses.
- This algorithm can be used on patient who have one or a combination of the following diseases: atelectasis, heart enlargement, standardization, edema, effusion, emphysema, fibrosis, hernia, infiltration, mass, Creed, pleura thickening and pneumothorax.

Device Limitations:

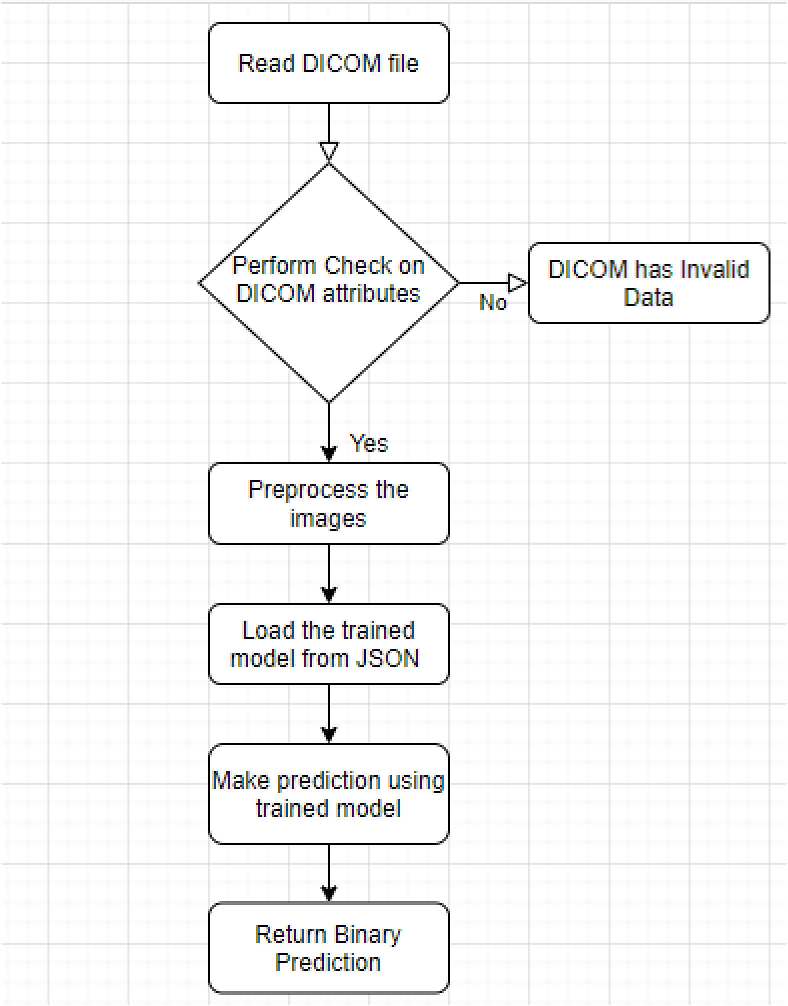
- Diagnosis can be made on a computer with at least 2-cores CPU and 8 GB RAM.
- Using GPU will faster the performance of algorithm.
- The algorithm's performance was measured on Intel(R) Core(TM)i3 @ 2.40GHz CPU with 8 GB RAM

Clinical Impact of Performance:

- There is a tradeoff between precision and recall. When the performance of the algorithm was evaluted using various metrics, we found that the algorithms has very high recall rate and low precision rate. Hence we can say that the algorithm is very good at predicting positive cases i.e this algorithm has high sensitivity. This fact can be used to rule out diseased patients. If for a patient the algorithm predicts negative, it is very likely that the patient does not have a disease because the has low false negatives. Hence can be successfully used for screening tests.

2. Algorithm Design and Function

Insert Algorithm Flowchart


Read DICOM file
Perform Check on DICOM attributes
No
DICOM has Invalid Data
Yes
Preprocess the images
Load the trained model from JSON
Make prediction using trained model
Return Binary Prediction

-

DICOM Checking Steps:

- While creating DICOM wrapper following checks were performed: -
  - Body part xxamined must be CHEST
  - Modality must be DX
  - Patient Position should be either AP or PA

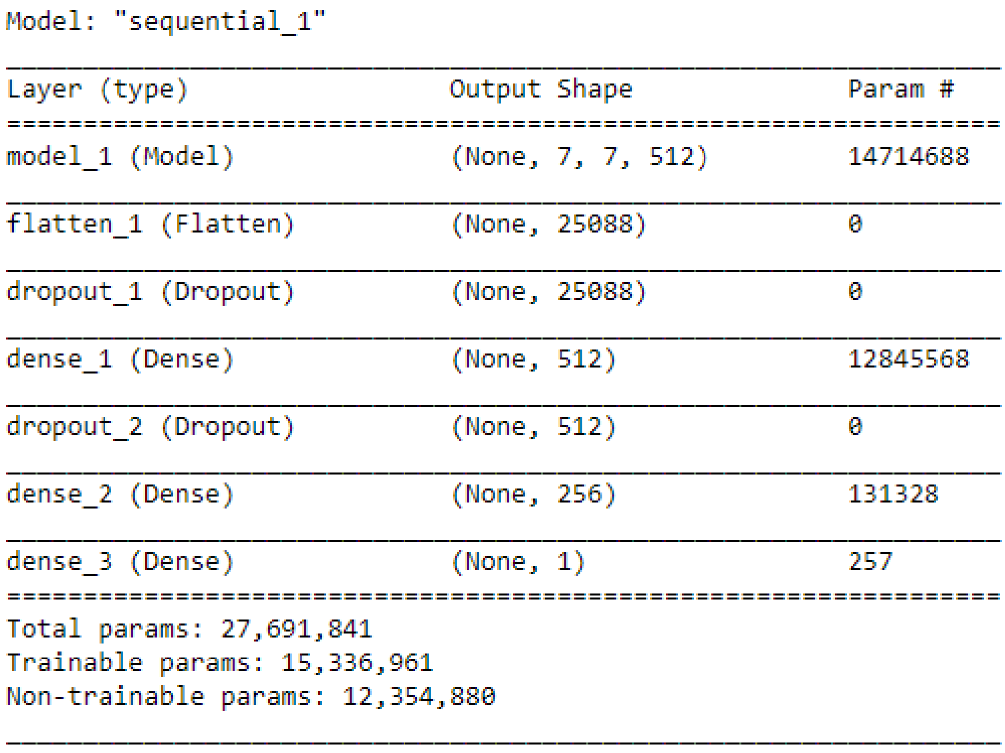
Preprocessing Steps:

- Following steps are performed during preprocessing
  - Images are converted to grayscale.
  - Pixels are normalized.
  - Images are reshaped to 224 x 224 to confrom with the input shape for VGG16 architecture.

CNN Architecture:

- The algorithm uses pre-trained VGG16 Neural Network, where only the last block of convolution and pooling layers were re-trained, with additional 2 blocks of fully Connected and dropout layers.
- The network output is a single probability value for binary classification.

- Here we have illustrated the model architecture in details (model_1 is VGG16 with only last layer trainable): -



```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
model_1 (Model)              (None, 7, 7, 512)         14714688  
_________________________________________________________________
flatten_1 (Flatten)          (None, 25088)             0         
_________________________________________________________________
dropout_1 (Dropout)          (None, 25088)             0         
_________________________________________________________________
dense_1 (Dense)              (None, 512)               12845568  
_________________________________________________________________
dropout_2 (Dropout)          (None, 512)               0         
_________________________________________________________________
dense_2 (Dense)              (None, 256)               131328    
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 257       
=================================================================
Total params: 27,691,841
Trainable params: 15,336,961
Non-trainable params: 12,354,880
_________________________________________________________________
```

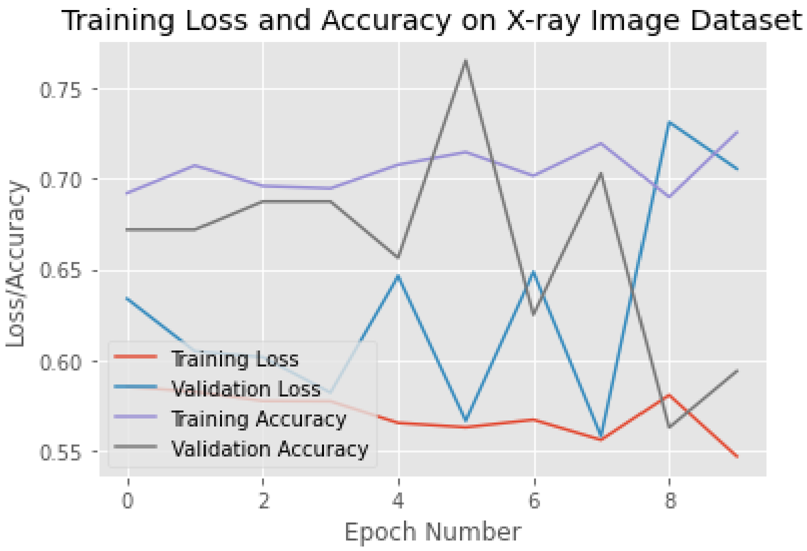
- Here is the Visual view of my model generated using visualkeras.

3. Algorithm Training

Parameters:

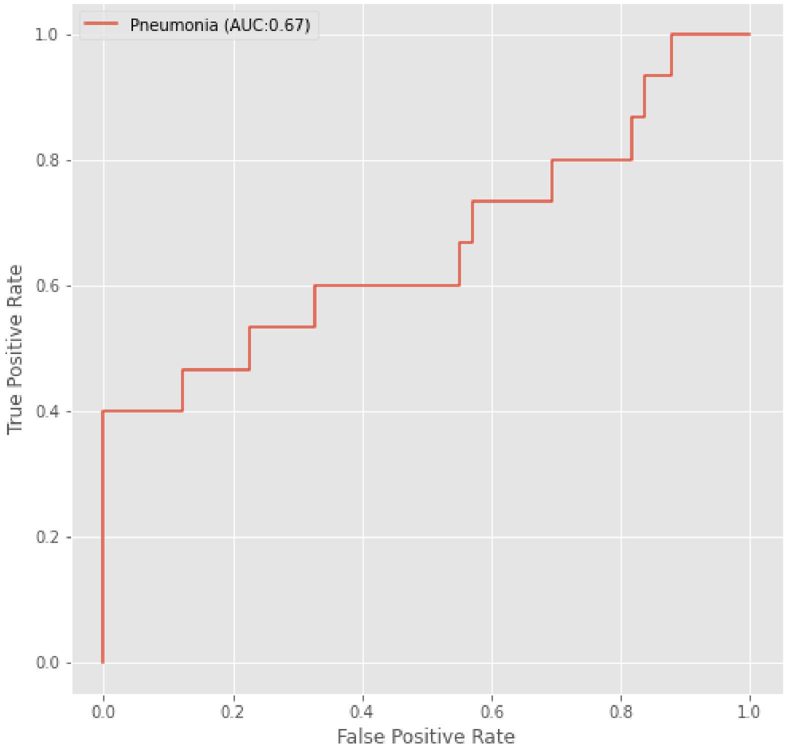
- Types of augmentation used during training
  - Augmentation on Training Set
    - rescaling of image pixels.
    - horizontal_flip.
    - width_shift_range of 0.1.
    - height_shift_range of 0.1.
    - shear_range of 0.1.
    - zoom_range of 0.2.
    - rotation_range of 15.
  - Augmentation on Validation Set
    - rescaling of image pixels.
- Batch size
  - Here I am using batch of 64 images.

- Optimizer learning rate
    - Here I am using learning rate of 0.0001 or (1e-4)
- Layers of pre-existing architecture that were frozen
    - input_4 (InputLayer)
    - block1_conv1 (Conv2D)
    - block1_conv2 (Conv2D)
    - block1_pool (MaxPooling2D)
    - block2_conv1 (Conv2D)
    - block2_conv2 (Conv2D)
    - block2_pool (MaxPooling2D)
    - block3_conv1 (Conv2D)
    - block3_conv2 (Conv2D)
    - block3_conv3 (Conv2D)
    - block3_pool (MaxPooling2D)
    - block4_conv1 (Conv2D)
    - block4_conv2 (Conv2D)
    - block4_conv3 (Conv2D)
    - block4_pool (MaxPooling2D)
    - block5_conv1 (Conv2D)
    - block5_conv2 (Conv2D)
- Layers of pre-existing architecture that were fine-tuned
    - block5_conv3 (Conv2D)
    - block5_pool (MaxPooling2D)
- Layers added to pre-existing architecture
    - flatten (Flatten)
    - dropout_1 (Dropout, 0.25)
    - dense_1 (Dense, 512)
    - dropout_2 (Dropout, 0.20)
    - dense_2 (Dense, 256)
    - dense_3 (Dense, 1)
- Algorithm training performance visualization
    - **Training Loss and Accuracy**



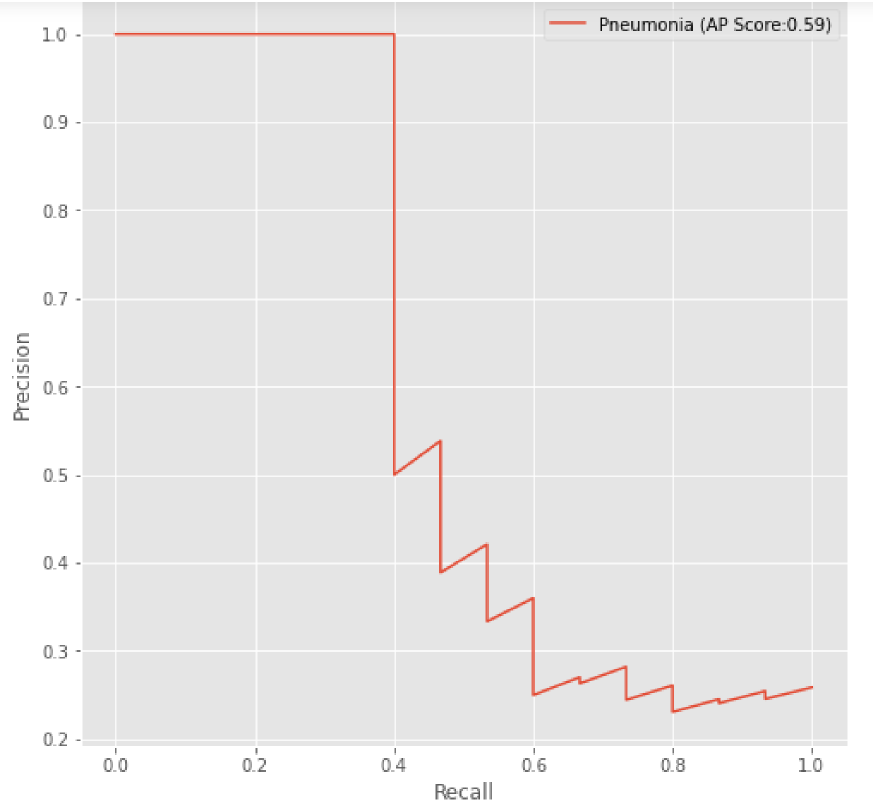
- ▪

- ◦ **ROC Curve**



- ▪

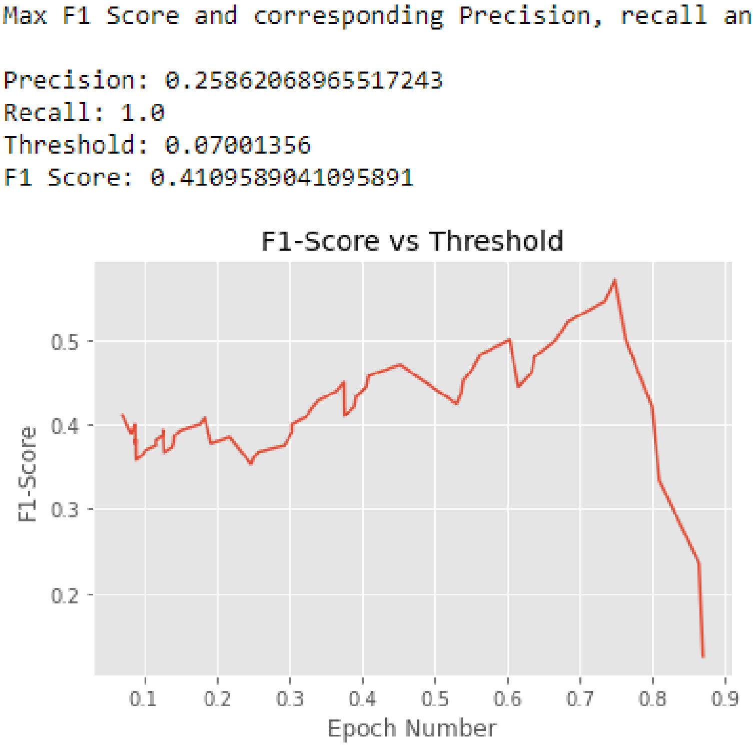
- ◦ **Precision Recall Curve**



▪

Final Threshold and Explanation:

- The final threshold is 0.07 because it gives the highest f1-score and can be seen clearly from below diagram.



```
Max F1 Score and corresponding Precision, recall and Threshold

Precision: 0.25862068965517243
Recall: 1.0
Threshold: 0.07001356
F1 Score: 0.4109589041095891
```

◦

- Here we need to make a trade off between recall and precision. High recall means the model will correctly classify all positive cases. High precision means the model will accuratly classify positive cases, which means when the model classifies an image as a positive case, the image will probably be a positive case. However, If we get a model with high recall, we will end up with many cases classified as positive. On the other hand, if we get a high precision model, we will end up missing positive cases.

4. Databases

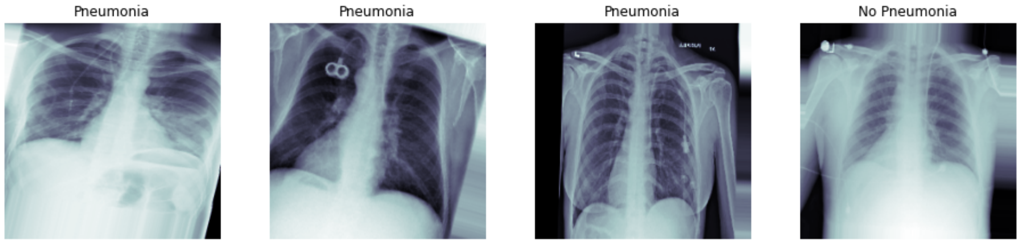
(For the below, include visualizations as they are useful and relevant)

NOTE: The dataset is obtained from Kaggle. The dataset contains 112,120 chest x-ray images with 1024x1024 resolution. It contains 14 diseases: atelectasis, heart enlargement, standardization, edema, effusion, emphysema, fibrosis, hernia, infiltration, mass, Creed, pleura thickening, pneumothorax, and pneumonia.

Description of Training Dataset:

- Training dataset consisted of 2290 chest xray images, with a 50/50 split between positive and negative cases of pneumonia.

- Here are some example images: -



Description of Validation Dataset:

- Validation dataset consisted of 1430 chest xray images, with 20/80 split between positive and negative cases, which more reflects the occurence of pneumonia in the real world.

5. Ground Truth

- The **ground truth** is obtained using Natural Language Processing (NLP) approach to mine the radiologist reports. Since, these labels were obtained using NLP, which is expected to be accurate enough for > 90% cases, still there might be some errenous labels.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:

- For the FDA Validation Dataset following population sample is suitable:
  - The sample should be taken from men and women aged 1 to 75 years.
  - The sample can not include people with previous comorbid thoracic pathologies.
  - X-rays should be for chest only with DX modality.

Ground Truth Acquisition Methodology:

- For obtaning ground truth we use Silver Standard approach, which is widely followed in the clinincal setting. What we can do is, we can hire few 3-4 radiologists and let them validate each of the X-Ray's and then final answer can be computed by weighted voting, by taking into acoount each of the radiologists experiences.

Algorithm Performance Standard:

- The algorithm's performance can be measured by calculating F1 score against 'silver standard' ground truth as described above. The algorithm's F1 score should exceed 0.387 which is an average F1 score taken over three human radiologists, as given in CheXNet.

- A 95% confidence interval given in the paper for average F1 score of 0.387 is (0.330, 0.442), so algorithm's 2.5% and 97.5% precentiles should also be calculated to get 95% confidence interval. This interval, when subtracted the interval above for the average, should not contain 0, which will indicate statistical significance of its improvement of the average F1 score. The same method for assessing statistical significance is presented in the above paper.